**Project Report**

IST652 Ava Nelson

**Data & Source**

The data for this assignment was taken from the Kaggle dataset, “Used Car Dataset” (https://www.kaggle.com/austinreese/craigslist-carstrucks-data/data). This dataset gathered over 430,000 used car listings from craigslist.com from across the country.

From the source dataset, I pulled 26,019 observations of selected columns to use for the following analysis. The observations were the “coupe” and “convertible” class vehicles. This brought the dataset from 430,000 observations of 25 columns to 26,019 observations using 14 columns. The trimmed dataset is “vehicles\_trimmed\_coupe.csv”.

This is a solo project.

**Data Exploration & Pre-Processing**

Data exploration began with importing the csv to Jupyter and verifying the data types of each variable. Python interpreted all number columns as float type, and other variables as string. As factors such as model year and odometer are not suited for float type, these columns were converted to integer.

A data type dictionary was built to apply against the imported dataset, to ensure that the data types will remain consistent each time the program is run.

Many NaN values were observed in the dataset. Due to the quantity of them across variables, it was elected that the values remain and only be overlooked on individual computations.

It is speculated that due to the platform the data was collected from (craigslist), that quality of the information given by the seller is inconstant. This can be seen in instances such as the Model column for index 4, which contains “Honda-Nissan-Kia-Ford-Hyundai-VW”. None of these descriptors are models of vehicles, so it may be inferred as user error.

To normalize the vehicle models for further analysis, the dataset was run against a dictionary of vehicle makes/models. (source: <http://tiresaddict.com/help/databases/cars/>).

For vehicle listings, text processing was done to remove any caps or punctuation that may be present, before looking to see if a model from the dictionary was textually present in the vehicle listing. If it was the string of the cell was replaced with the normalized value, while any listings without a dictionary model present was discarded. Due to limitations of the dictionary this process was not perfect and removed the bulk of listings—bringing the dataset to 2,880 observations.

These errors were also kept, as they should not impact the intent of the resulting program, and there is no good way to interpret what should correctly fill these cells.

.

**Question**

The question selected for this assignment is as follows:

**Question:** What caliber of sports car can be purchased from Craigslist? Of these, what common metrics present the most influence over price, and which vehicle models retain monetary value most consistently?

**Program Description**

First, the program calls pandas, numpy, intertools, seaborn, and matplotlib into python.

As stated previously, the program reads in the data and runs pre-processing to normalize vehicle models for further analysis. The normalized dataframe can be seen as csv under “vehicles\_final.csv”.

The program then writes a function to express statistics on present vehicle manufacturers. Statistics include oldest model year, newest model year, cheapest vehicle, and so on. It also generates this table for the vehicle model most occurring for each manufacturer.

In a fresh Notebook, the normalized data is run through exploratory data visualization with the graphics shown as output. Graphics display the distribution of listings across manufacturers, distribution of listings by year and price, and distribution of listings filtered by year, price, and manufacturer.

Exploratory visualization was also run against the dataset, observing the correlation of variables to one another.

A Model Search was created where a user may input a vehicle model to return a profile for the vehicle in question from the craigslist data. The profile returns the name of the vehicle, number of listings, and average cost for the vehicle. It also returns these metrics for common sports car purchase considerations (body color and engine). The Model Search also returns a comparison graphic showing the distribution of this vehicle across model year, price, and odometer. The intent for this function was to provide useful metrics to compare a vehicle against others like it in the marketplace. These metrics provide a useful gauge for what pricing should be considered fair for a vehicle.

Some Model Searches were made to showcase a variety of vehicle models.

**Output Files**

The output files are the CSVs containing the summary comparison information as outlined above (vehicles\_final.csv), along with the grouped vehicle csv (vehicles\_grouped.csv). Also see (BrandStats.csv) and (ModelStats.csv) for further summation of data.

Additional files are graphics present in the Jupyter Notebook code/output file.

**Conclusion**

There were a wide variety of coupe vehicles present on craigslist. Of these, old classics were mixed in with old daily drivers and rare, expensive exotics. It is worth mentioning that my express vehicles of interest were not highly present on the dictionary or craigslist, so the focus of this analysis adjusted accordingly.

Of vehicle manufacturers, Chevrolet possessed the highest variety of vehicle models in their line up on Craigslist. Audi, Ford, and Toyota both had high variety as well. In regards to Chevrolet, Ford, and Toyota, the presence of 2 door sports versions of their sedans accounted for a good degree of this representation outside of their sports car line ups.

Most vehicles on Craigslist are under 20 years in age. Vehicle price tends to be most expensive for vehicles younger than 10 years in age and for vehicles between the model years 1960-1980. The “sweet spot” for inexpensive vehicles is in the 1990s model years. Based on this, if a user wanted the cheapest vehicle possible it is recommended they target vehicles built from 1990-2000. Judging by the nature of the price increases in vehicles from 1960-1980, it is probable that a smart 1990s era vehicle purchase could result in a net gain of value over time.

Toyotas are overall less expensive than other manufacturers. Audi contains the most expensive vehicles listed, with Ferrari not far behind. Unsurprisingly, the upper echelons on price for used vehicles is dominated by luxury brands.

One interesting observation is the introduction of Asian and non-domestic auto manufacturers into the American mainstream as seen through these listings. There is a clear indication around 1970 where many overseas vehicles begin to appear.

For the most inexpensive vehicle models, the Ford Mustang (mean price $9,998), Honda Civic ($6,470), Honda CRZ ($6,841), Hyundai Veloster ($8,552), and Toyota Camry ($5,832) all held an average value under $10,000.

Of these, on average, the Hyundai Veloster had the most recent models (2014 model year) and the fewest overall average mileage (60,852 miles). This indicates the Veloster as being a very good buy for the quality of used car available for the price.

Looking at the Model Search Reports, for the Veloster the most prevalent body color available on this marketplace is black. This color appears to be less desirable, as the average cost for a black Veloster is $500 less than the average overall price. In comparison, a black civic appears to be more desirable a color than other civics, with the average price for a black civic $500 more than the overall.

Based on this data analysis, it appears that the Hyundai Veloster is the optimal for a low cost, “new”, fun sports car.

In answering the question proposal,it appears that body color has moderate influence over vehicle price. Model Year and Model itself has the greatest influence over price. The Hyundai Veloster appears to retain monetary value quite well.